Notice of Allowability	Application No.	Applicant(s)	
	10/605,630	FRANCIOSA ET AL.	
	Examiner	Art Unit	
	Usmaan Saeed	2166	

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address--

All claims being allowable, PROSECUTION ON THE MERITS IS (OR REMAINS) CLOSED in this application. If not included herewith (or previously mailed), a Notice of Allowance (PTOL-85) or other appropriate communication will be mailed in due course. **THIS NOTICE OF ALLOWABILITY IS NOT A GRANT OF PATENT RIGHTS.** This application is subject to withdrawal from issue at the initiative of the Office or upon petition by the applicant. See 37 CFR 1.313 and MPEP 1308.

1. ☒ This communication is responsive to *the amendment dated 11/05/2007*.

2. ☒ The allowed claim(s) is/are *1, 3-4, 6-7, and 9-22 (renumbered as 1-19)* .

3. ☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).
   a) ☐ All b) ☐ Some* c) ☐ None of the:
      1. ☐ Certified copies of the priority documents have been received.
      2. ☐ Certified copies of the priority documents have been received in Application No. _____ .
      3. ☐ Copies of the certified copies of the priority documents have been received in this national stage application from the International Bureau (PCT Rule 17.2(a)).

   * Certified copies not received: _____.

Applicant has THREE MONTHS FROM THE "MAILING DATE" of this communication to file a reply complying with the requirements noted below. Failure to timely comply will result in ABANDONMENT of this application.
THIS THREE-MONTH PERIOD IS NOT EXTENDABLE.

4. ☐ A SUBSTITUTE OATH OR DECLARATION must be submitted. Note the attached EXAMINER'S AMENDMENT or NOTICE OF INFORMAL PATENT APPLICATION (PTO-152) which gives reason(s) why the oath or declaration is deficient.

5. ☐ CORRECTED DRAWINGS (as "replacement sheets") must be submitted.
   (a) ☐ including changes required by the Notice of Draftsperson's Patent Drawing Review (PTO-948) attached
      1) ☐ hereto or 2) ☐ to Paper No./Mail Date _____.
   (b) ☐ including changes required by the attached Examiner's Amendment / Comment or in the Office action of Paper No./Mail Date _____.

   Identifying indicia such as the application number (see 37 CFR 1.84(c)) should be written on the drawings in the front (not the back) of each sheet. Replacement sheet(s) should be labeled as such in the header according to 37 CFR 1.121(d).

6. ☐ DEPOSIT OF and/or INFORMATION about the deposit of BIOLOGICAL MATERIAL must be submitted. Note the attached Examiner's comment regarding REQUIREMENT FOR THE DEPOSIT OF BIOLOGICAL MATERIAL.

Attachment(s)

1. ☒ Notice of References Cited (PTO-892)
2. ☐ Notice of Draftperson's Patent Drawing Review (PTO-948)
3. ☐ Information Disclosure Statements (PTO/SB/08), Paper No./Mail Date _____
4. ☐ Examiner's Comment Regarding Requirement for Deposit of Biological Material
5. ☐ Notice of Informal Patent Application
6. ☒ Interview Summary (PTO-413), Paper No./Mail Date *1/17/08* .
7. ☒ Examiner's Amendment/Comment
8. ☒ Examiner's Statement of Reasons for Allowance
9. ☒ Other *See Continuation Sheet*.

HOSAIN ALAM
SUPERVISORY PATENT EXAMINER

Continuation of Attachment(s) 9. Other: Copy of Applicant's email for Examiners proposed amendment.

DETAILED ACTION

1. This communication is in response to the amendment filed on 11/05/2007.

After thorough search and examination of the present application and in light of the prior art made of record, claims 1, 3-4, 6-7, and 9-22 (renumbered as 1-19) are allowed.

Claims 2, 5, and 8 have been cancelled.

EXAMINER'S AMENDMENT

2. An examiner's amendment to the record appears below. Should the changes and/or additions be unacceptable to applicant, an amendment may be filed as provided by 37 CFR 1.312. To ensure consideration of such an amendment, it MUST be submitted no later than the payment of the issue fee.

Authorization for this examiner's amendment was given in a telephone interview/email with Attorney, Kevin M. Dunn, Registration No. 52,842 on January 17, 2008.

A copy of the attorneys email regarding the proposed examiner's amendment is also attached.

Please amend the claims, which were filed on 11/05/2007 with the new version as follows:

1. (Currently amended) A computer implemented method for identifying output documents similar to an input document, comprising:

(a) receiving the input document that includes textual content;

(b) performing optical character recognition on the textual content to identify text;

(c) analyzing the text and the textual content to identify keywords, wherein a predefined number of keywords is identified from a first list of rated keywords extracted from the input document;

(d) creating a list of best keywords wherein for each keyword remaining in the first list of keywords performing the steps,

(1) identifying the keyword in one or more domain specific dictionaries of words and phrases in which they are used;

(2) identifying combinations of keywords in the list of keywords that satisfy the longest phrase;

(3) determining the frequency of occurrence in the input document of the identified keywords and phrases identified in the one or more domain specific dictionaries;

(4) setting the linguistic frequency of occurrence of the keywords and phrases to a predefined value; and

(e) defining a list of best keywords, wherein the list of best keywords has a rating greater than other keywords in the first list of keywords except for keywords belonging to a domain specific dictionary of words and having no measurable linguistic frequency;

(f) formulating a query using the list of best keywords;

(g) performing the query to assemble a first set of output documents;

(h) identifying lists of keywords for each output document in the first set of documents by tokenizing the keywords at one or more predefined word boundaries while maintaining order of the sequence of the input text and translating the keywords into one or more languages;

(i) computing a measure of similarity between the input document and each output document in the first set of documents; and

(j) defining a second set of documents with each document in the first set of documents for which its computed measure of similarity with the input document is greater than a predetermined threshold value; wherein the list of best keywords has a maximum number of keywords less than the number of keywords in the list of best keywords that are identified as belonging to a domain specific dictionary of words and having no measurable linguistic frequency, each document in the second set of documents is identified as being one of a match, a revision, and a relation of the input document, wherein the query is repeated until a predetermined number of results are obtained or the query is terminated;

(k) if the second set of documents includes a matching document but no similar documents repeating (a)-(j) using the matching document to identify similar documents, wherein if one or more documents is related to a copyright registered document, the one or more documents is rights limited; and

(l) delivering each document in the second set of documents to one or more predetermined output devices, wherein the collection of documents is set forth in a list serialized in XML that contains for each document found: its location on a network,

original representation, unformatted representation, service results, metadata, distance measurement, type of document found according to desired quality, and error status.

2. (Cancelled)

3. (Currently amended) The method according to claim 1, further comprising:

(m) if the second set of document contains an insufficient number of output documents, performing query reduction by removing at least one keyword in the list of best keywords that is not the keyword that is identified as belonging to a domain specific dictionary and having no measurable linguistic frequency.

4. (Currently amended) The method according to claim 3, further comprising if after performing (m) the second set of document contains an insufficient number of output documents, performing:

(n): replacing the list of best keywords using keywords having a rating greater than other keywords in the first list of rated keywords; and repeating (b)-(l).

5. (Cancelled)

6. (Previously presented) The method according to claim 4, performing (n) when textual content in the input document is identified using OCR or a portion of the input document matches the output document.

7. (Previously presented) The method according to claim 1, wherein the predefined number of keywords identified from the first list of rated keywords is five.

8. (Cancelled)

9. (Original) The method according to claim 1, further comprising:

recording a digital image representation of the input document;

performing OCR on the digital image representation to identify text;

analyzing the text to identify keywords.

10. (Previously presented) The method according to claim 1, further comprising:

(o) extracting from the input document the first list of keywords;

(p) determining if each keyword in the first list of keywords exists in a domain specific dictionary of words;

(q) for each keyword in the first list of keywords, determining its frequency of occurrence in the input document, also referred to as its term frequency;

(r) for each keyword identified at (k) that exists in the domain specific dictionary of words, assigning each keyword its linguistic frequency if one exists from a database of linguistic frequencies defined using a collection of documents, and assigning its

linguistic frequency to a predefined small value if one does not exist in the database of linguistic frequencies;

(s) for each keyword that was not identified in the domain specific dictionary of words at (h), assigning each keyword its linguistic frequency if one exists in the database of linguistic frequencies; and

(t) for each keyword in the first list of keywords to which a term frequency and a linguistic frequency are assigned, computing a rating corresponding to its importance in the input document that is a function of its frequency of occurrence in the input document and its frequency of occurrence in the collection of documents.

11. (Previously presented) The method according to claim 10, for each keyword that was not identified in the domain specific dictionary of words at (p) and that was not assigned at (r) a linguistic frequency from the database of linguistic frequencies, assigning each that matches a regular expression from a set of regular expressions a predefined rating.

12. (Previously presented) The method according to claim 11, further comprising, for each keyword in the first list of keywords, modifying the term frequency of keywords determined at (q) to a predefined maximum.

13. (Original) The method according to claim 12, wherein keywords include phrases of keywords.

14. (Original) The method according to claim 11, wherein the rating is a weight computed using the following equation: $W_{t,d} = F_{t,d} * \log(N / F_t)$, where:

$W_{t,d}$: the weight of term t in document d;
$F_{t,d}$: the frequency occurrence of term t in document d;
N: the number of documents in the collection of documents;
F_t: the document linguistic frequency of term t in the collection of documents.

15. (Original) The method according to claim 11, wherein keywords that do not match a regular expression from the set of regular expressions are removed from the first list of keywords.

16. (Currently amended) A computer implemented method for computing ratings of keywords extracted from an input document, comprising:

(a) determining if each keyword in the list of keywords exists in a domain specific dictionary of words by tokenizing the keywords at one or more predefined word boundaries while maintaining order of the sequence of the input text and translating the keywords into one or more languages;

(b) determining a frequency of occurrence in the input document for each keyword in the list of keywords, also referred to as its term frequency;

(c) for each keyword identified at (a) that exists in the domain specific dictionary of words, assigning each keyword its linguistic frequency if one exists from a database of linguistic frequencies defined using a collection of documents, and assigning its

linguistic frequency to a predefined small value if one does not exist in the database of linguistic frequencies;

(d) for each keyword that was not identified in the domain specific dictionary of words at (a), assigning each keyword its linguistic frequency if one exists in the database of linguistic frequencies; and

(e) for each keyword in the list of keywords to which a term frequency and a linguistic frequency are assigned, computing a rating corresponding to its importance in the input document that is a function of its frequency of occurrence in the input document and its frequency of occurrence in the collection of documents, wherein a query reduction is performed by removing at least one keyword in the list of best keywords that is identified as belonging to a domain specific dictionary and having no measurable linguistic frequency if an insufficient number of results are obtained from the list of keywords, wherein the query is repeated until a predetermined number of results are obtained or the query is terminated;, ~~wherein if one or more documents is a copy of a known copyright registered document, the one or more documents is rights limited;~~ ~~and~~

(f) defining a list of best keywords wherein the list of best keywords have a rating greater than other keywords in the list of keywords except for keywords belonging to a domain specific dictionary of words and having no measurable linguistic frequency by tokenizing the keywords at one or more predefined word boundaries while maintaining order of the sequence of the input text and translating the keywords into one or more languages;

(g) formulating a query using the list of best keywords;

(h) performing the query to assemble a first set of output documents;

(i) identifying lists of keywords for each output document in the first set of documents;

(j) computing a measure of similarity between the input document and each output document in the first set of documents;

(k) defining a second set of documents with each document in the first set of documents for which its computed measure of similarity with the input document is greater than a predetermined threshold value; wherein the list of best keywords has a maximum number of keywords less than the number of keywords in the list of best keywords that are identified as belonging to a domain specific dictionary of words and having no measurable linguistic frequency, each document in the second set of documents is identified as being one of a match, a revision, and a relation of the input document; and

(l) delivering each document in the collection of documents to a predetermined output device, wherein the collection of documents is set forth in a list serialized in XML that contains for each document found: its location on a network, original representation, unformatted representation, service results, metadata, distance measurement, type of document found according to desired quality, and error status.

17. (Original) The method according to claim 16, wherein the keywords in the list of keywords are used to carry out one of language identification, indexing,

categorization, clustering, searching, translating, storing, duplicate detection, and filtering.

18. (Currently amended) A computer implemented system for identifying output documents similar to an input document, comprising: a memory for storing the output documents and the input document and processing instructions of the system; and a processor coupled to the memory for executing the processing instructions of the system; the processor in executing the processing instructions:

(a) identifying a predefined number of keywords from a first list of rated keywords extracted from the input document;

(b) creating a list of best keywords wherein for each keyword remaining in the first list of keywords performing the steps,

(1) identifying the keyword in one or more domain specific dictionaries of words and phrases in which they are used;

(2) identifying combinations of keywords in the list of keywords that satisfy the longest phrase;

(3) determining the frequency of occurrence in the input document of the identified keywords and phrases identified in the one or more domain specific dictionaries;

(4) setting the linguistic frequency of occurrence of the keywords and phrases to a predefined value; and

(c) defining a list of best keywords wherein the list of best keywords have a rating greater than other keywords in the first list of keywords except for keywords belonging to a domain specific dictionary of words and having no measurable linguistic frequency by tokenizing the keywords at one or more predefined word boundaries while maintaining order of the sequence of the input text and translating the keywords into one or more languages;

(d) formulating a query using the list of best keywords;

(e) performing the query to assemble a first set of output documents;

(f) identifying lists of keywords for each output document in the first set of documents;

(g) computing a measure of similarity between the input document and each output document in the first set of documents;

(h) defining a second set of documents with each document in the first set of documents for which its computed measure of similarity with the input document is greater than a predetermined threshold value; wherein the list of best keywords has a maximum number of keywords less than the number of keywords in the list of best keywords that are identified as belonging to a domain specific dictionary of words and having no measurable linguistic frequency; and

(i) if the second set of document contains an insufficient number of output documents, performing query reduction by removing at least one keyword in the list of best keywords that is not the keyword that is identified as belonging to a domain specific

dictionary and having no measurable linguistic frequency, wherein the query is repeated until a predetermined number of results are obtained or the query is terminated;

(j) if the second set of documents includes a matching document but no similar documents repeating (a)-(i) using the matching document to identify similar documents ~~,~~

~~wherein if one or more documents is a copy of a known copyright registered document, the one or more documents is rights limited~~; and

(k) delivering each document in the second set of documents to a predetermined output device, wherein the collection of documents is set forth in a list serialized in XML that contains for each document found: its location on a network, original representation, unformatted representation, service results, metadata, distance measurement, type of document found according to desired quality, and error status.

19. (Currently amended) The system according to claim 18, wherein the processor in executing the processing instructions further comprises:

~~(k)~~ (l) extracting from the input document the first list of keywords;

~~(l)~~ (m) determining if each keyword in the first list of keywords exists in a domain specific dictionary of words;

~~(m)~~ (n) for each keyword in the first list of keywords, means for determining its frequency of occurrence in the input document, also referred to as its term frequency;

~~(n)~~ (o) for each keyword identified at (m) that exists in the domain specific dictionary of words, means for assigning each keyword its linguistic frequency if one exists from a database of linguistic frequencies defined using a collection of documents,

and assigning its linguistic frequency to a predefined small value if one does not exist in the database of linguistic frequencies;

~~(o)~~ (p) for each keyword that was not identified in the domain specific dictionary of words at (l), means for assigning each keyword its linguistic frequency if one exists in the database of linguistic frequencies; and

~~(p)~~ (q) for each keyword in the first list of keywords to which a term frequency and a linguistic frequency are assigned, means for computing a rating corresponding to its importance in the input document that is a function of its frequency of occurrence in the input document and its frequency of occurrence in the collection of documents.

20. (Currently amended) An article of manufacture for identifying output documents similar to an input document, the article of manufacture comprising computer usable storage media including computer readable instructions embedded therein that causes a computer to perform a method, wherein the method comprises:

(a) identifying a predefined number of keywords from a first list of rated keywords extracted from the input document to define a list of best keywords; the list of best keywords having a rating greater than other keywords in the first list of keywords except for keywords belonging to a domain specific dictionary of words and having no measurable linguistic frequency, wherein the keywords are tokenized at one or more predefined word boundaries while maintaining order of the sequence of the input text and translating the keywords into one or more languages;

(b) creating a list of best keywords wherein for each keyword remaining in the first list of keywords performing the steps,

(1) identifying the keyword in one or more domain specific dictionaries of words and phrases in which they are used;

(2) identifying combinations of keywords in the list of keywords that satisfy the longest phrase;

(3) determining the frequency of occurrence in the input document of the identified keywords and phrases identified in the one or more domain specific dictionaries;

(4) setting the linguistic frequency of occurrence of the keywords and phrases to a predefined value; and

(c) defining a list of best keywords wherein the list of best keywords have a rating greater than other keywords in the first list of keywords except for keywords belonging to a domain specific dictionary of words and having no measurable linguistic frequency by tokenizing the keywords at one or more predefined word boundaries while maintaining order of the sequence of the input text and translating the keywords into one or more languages;

(d) formulating a query using the list of best keywords;

(e) performing the query to assemble a first set of output documents;

(f) identifying lists of keywords for each output document in the first set of documents;

(g) computing a measure of similarity between the input document and each output document in the first set of documents;

(h) defining a second set of documents with each document in the first set of documents for which its computed measure of similarity with the input document is greater than a predetermined threshold value; wherein the list of best keywords has a maximum number of keywords less than the number of keywords in the list of best keywords that are identified as belonging to a domain specific dictionary of words and having no measurable linguistic frequency, each document in the second set of documents is identified as being one of a match, a revision, and a relation of the input document; and

(i) if the second set of document contains an insufficient number of output documents, performing query reduction by removing at least one keyword in the list of best keywords that is not the keyword that is identified as belonging to a domain specific dictionary and having no measurable linguistic frequency, wherein the query is repeated until a predetermined number of results are obtained or the query is terminated;

(j) if the second set of documents includes a matching document but no similar documents repeating (a)-(i) using the matching document to identify similar documents; and

(k) delivering each document in the second set of documents to a predetermined output device, wherein the collection of documents is set forth in a list serialized in XML that contains for each document found: its location on a network, original

representation, unformatted representation, service results, metadata, distance measurement, type of document found according to desired quality, and error status.

21. (Previously presented) The system according to claim 18, further comprising if after performing (i) the second set of document contains an insufficient number of output documents, performing:

(l) replacing the list of best keywords using keywords having a rating greater than other keywords in the first list of rated keywords; and repeating (b)-(k).

22. (Previously presented) The system according to claim 18, wherein for each keyword that was not identified in the domain specific dictionary of words at (h) and that was not assigned at (i) a linguistic frequency from the database of linguistic frequencies, assigning each that matches a regular expression from a set of regular expressions a predefined rating, wherein the rating is a weight computed using the following equation: $W_{t,d} = F_{t,d} * \log(N / F_t)$, where:

$W_{t,d}$: the weight of term t in document d;

$F_{t,d}$: the frequency occurrence of term t in document d;

N: the number of documents in the collection of documents;

F_t: the document linguistic frequency of term t in the collection of documents.

Reason for Allowance

3. The prior art made of record does not teach or fairly suggest the combination of elements, as recited in independent claims 1, 16, 18, and 20.

More specifically, the prior art of records does not specifically suggest the combination of "identifying combinations of keywords in the list of keywords that satisfy the longest phrase; determining the frequency of occurrence in the input document of the identified keywords and phrases identified in the one or more domain specific dictionaries; wherein the keywords are tokenized at one or more predefined word boundaries while maintaining order of the sequence of the input text and translating the keywords into one or more languages; and delivering each document in the second set of documents to a predetermined output device, wherein the collection of documents is set forth in a list serialized in XML that contains for each document found: its location on a network, original representation, unformatted representation, service results, metadata, distance measurement, type of document found according to desired quality, and error status" in combination with all the other limitations in the independent claims 1, 16, 18, and 20.

These features together with other limitations of the independent claims are novel and non-obvious over the prior art of record. The dependent claims 3-4, 6-7, 9-15, 17, 19, and 21-22 being definite, enabled by the specification, and further limiting to the independent claims, are also allowable.

Any comments considered necessary by applicant must be submitted no later than the payment of the issue fee and, to avoid processing delays, should preferably

accompany the issue fee. Such submissions should be clearly labeled "Comments on Statement of Reasons for Allowance."

Contact Information

4. Any inquiry concerning this communication or earlier communications from the examiner should be directed to Usmaan Saeed whose telephone number is (571)272-4046. The examiner can normally be reached on M-F 8-5.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Hosain Alam can be reached on (571)272-3978. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see http://pair-direct.uspto.gov. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free).

Usmaan Saeed
Patent Examiner
Art Unit: 2166

Hosain Alam
Supervisory Patent Examiner

US
January 18, 2008